图书情报工作
第64卷 第9期 2020年5月

在线社交网络的社区发现研究进展

■ 张海涛[1,2] 周红磊[1] 张鑫蕊[1] 孙彤[1]
[1] 吉林大学管理学院 长春 130022 [2]吉林大学信息资源研究中心 长春 130022

摘 要：[目的/意义]以在线社交网络为研究对象，通过文献梳理准确捕捉社区发现的发展态势及研究热点，探究如何在大规模社交网络中挖掘隐藏社区，具有理论和实践意义。[方法/过程]以中国知网(CNKI)数据库、Web of Science核心合集及相关国际会议文献作为数据来源，应用CiteSpace可视化分析工具从热点关键词、主题演化路径以及共被引文献等方面进行定量研究，并从社区发现方法、算法实现及应用实践3个维度对文献内容详细述评。[结果/结论]当前研究领域仍有广阔的发展空间，未来应注重算法的优化及创新、应用场景的区分和拓展以及融合跨学科知识、前沿技术方法的交叉研究。

关键词：在线社交网络 社区发现 动态社区演化 研究进展
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.09.016

1 引言

近年来，在线社交平台对个人学习生活、国家经济发展以及社会稳定带来重要影响，知乎、微博等社会化媒体拓宽了网民汲取知识、了解新闻动态的渠道。然而，在线社交网络平台全面融入人们生活的同时，虚假谣言信息盛行、网络推手和欺诈活动频繁出现，扰乱了正常的互联网秩序。方滨兴等[1]认为，在线社交网络是指信息网络上由群体集合及个体联系构成的社会性结构，包含网络群体、关系结构以及网络信息3个要素，呈现出人际传播与虚拟交互相渗透的特点。

研究发现，在线社交网络中存在与现实社会一样的社区结构，即整个社交网络由若干子社区构成[2]，发现这些潜在社区对研究信息传播等级、好友推荐以及网络舆情监管等方面具有现实意义。在学术研究方面，知识网络的结构演化也反映出学科研究主题的发展过程，社区发现方法被广泛应用于作者合著网络及学科知识流动等课题[3-4]，成为研究学者合作网络、学术主题演化的新视角。社区发现也被称为社区检测、社区识别以及社群发现等，M. E. J. Newman[5]认为社区发现是将整个网络结构依据网络节点划分成若干小组，使得组内节点连接较为稠密，组间节点连接较为稀疏，其中小组则为深度挖掘到的隐藏社区或子社区。简单而言，社区发现的涵义即在整个网络结构中发现存在某种关联性的子社区。但随研究逐步深入，在线社交网络中的社区划分依据有所扩展，可以借助兴趣网络、标签信息以及节点链接等刻画在线社交网络的社区结构，因此，概念仍属于不断更迭的范畴，至今没有明确定义。

基于上述逻辑，本文对相关文献进行定量分析，并根据研究热点将文献归纳为社区发现的方法、算法以及应用实践3个维度，分析提炼国内外该主题的发展态势及研究前沿，以期明晰未来研究的切入点，具体研究思路见图1。

2 在线社交网络的社区发现主题文献计量分析

运用可视化分析软件CiteSpace对主题文献进行计量分析，包括论文年代分布、文献关键词统计、主题发展脉络以及共被引文献分析4个方面。

2.1 文献来源

本文数据主要来自3个方面：中英文权威数据库、国际会议期刊以及在线学术社区和自媒体平台，具体来源分布如下：

(1)英文文献来源。基于Web of Science核心集，

作者简介：张海涛(ORCID:0000-0002-9421-8187)，教授，博士生导师，E-mail：zhtinfo@126.com；周红磊(ORCID:0000-0002-9732-8138)，硕士研究生；张鑫蕊(ORCID:0000-0001-9413-6109)，硕士研究生；孙彤(ORCID:0000-0002-0068-0275)，硕士研究生。

收稿日期：2019-11-12 修回日期：2020-02-07 本文起止页码：142-152 本文责任编辑：易飞

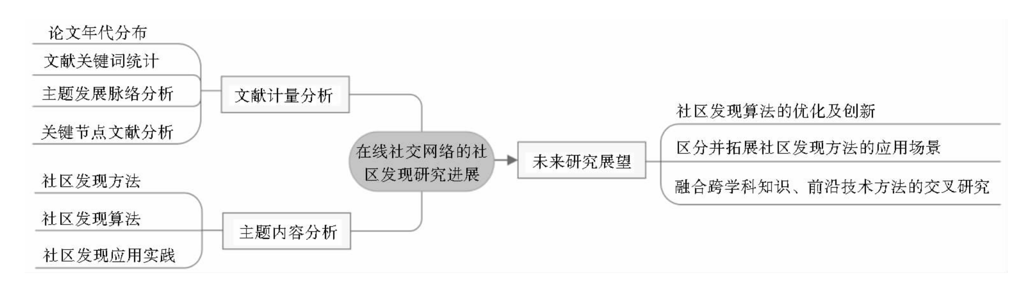


图1 本文研究逻辑

时间跨度为1990-2019年，检索式为：TS=("Online social network" and("Community discovery" or "community identif *" or" Community detect *"))，共获取885篇文献，根据主题和文摘内容，筛选得到460篇高度相关文献；在Elsevier以同样方式检索，获得补充文献17篇；另在文章撰写过程中，借鉴相关领域最新顶级期刊会议文献52篇。

(2)中文文献来源。主要基于CNKI、万方以及维普中文科技期刊3个数据库，检索式为：SU=('社交网络'+'社会化媒体平台')*('社区发现'+'社区检测'+'社区识别'+'社群发现')，共检索到439篇中文文献，经去重、筛选后得到402篇高度相关文献。

(3)在线学术社区及自媒体平台。借鉴当前较为科学的学术社区及自媒体平台如简书、今日头条、CSDN等平台，研究标签传播算法、Louvain等常用算法的实际效果及效率测评。

2.2 论文年代分布

国外于1999年首次针对社交网络中的社区结构进行探索，提出了大量开创性想法和基础理论。如图2所示，国内外发文量整体呈同步增长态势，2004年是国内首次研究该主题，主要是外文文献的综述、国外权威算法的借鉴及优化等；2008年，在线社交平台、娱乐性博客以及知识分享类学习社区快速发展，文献数量首次呈现大幅度增长态势；2010年至今，是该领域的研究热潮，涌现出大量的算法和方法研究，实际应用场景也得到丰富，主要涉及到通讯业、电子商务行业以及网络安全等领域。

2.3 文献关键词统计

国内关键词分布如图3所示，除社区发现及社交网络以外，复杂网络是最重要的节点，其出现频次达到51次。此外，词频较高的节点包括微博(40次)、重叠社区(29次)、标签传播(27次)、社区划分(20次)、社区结构(18次)、协同过滤(18次)、推荐系统(16次)、主题模型(11次)、模块度(11次)、矩阵分解(10次)以及好友推荐(10次)等，为下文主题内容分析提供了参考。

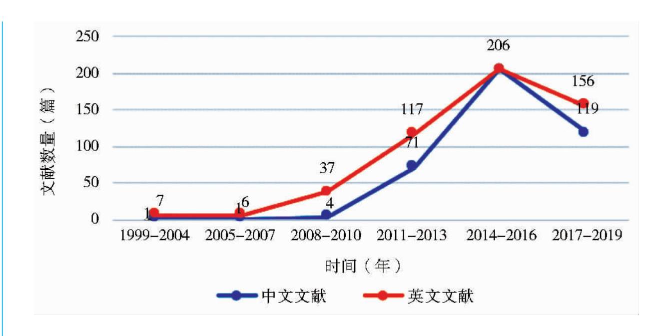


图2 论文年代分布



图3 国内社区发现研究热点

国外关键词分布如图4所示，与国内相似，除"社区发现"和"社交网络"节点外，出现频次较高的重要节点包括复杂网络(105次)、社交网络分析(78次)、模块度(57次)、算法(50次)、重叠社区(43次)、社区结构(27次)、图谱(27次)、集群(21次)、标签算法(17次)、中心度(16次)以及数据挖掘(12次)等。



图4 国外社区发现研究热点

2.4　主题发展脉络分析

本部分应用时区视图对研究主题演化路径做出阐释。国内“社区发现”研究始于 2004 年，杨楠等[6]对 web 社区发现技术做了述评。“复杂网络”主题从 2004 年一直延续至今，其中小世界性质、无标度性质以及网络聚类等特性得到充分研究。2005 – 2008 年，研究主题空白明显。2010 年开始，该领域得到广泛关注，社区发现方法在信息检索、推荐系统及用户分析等方面应用广泛，以微博为主的社会化媒体成为在线社交网络研究的重要平台，用户的倾向性分析、兴趣扩展、话题检测得到关注。网络社区演化、社区结构识别以及“动态社区”和“重叠社区”发现算法均为热点话题。如图 5 所示：

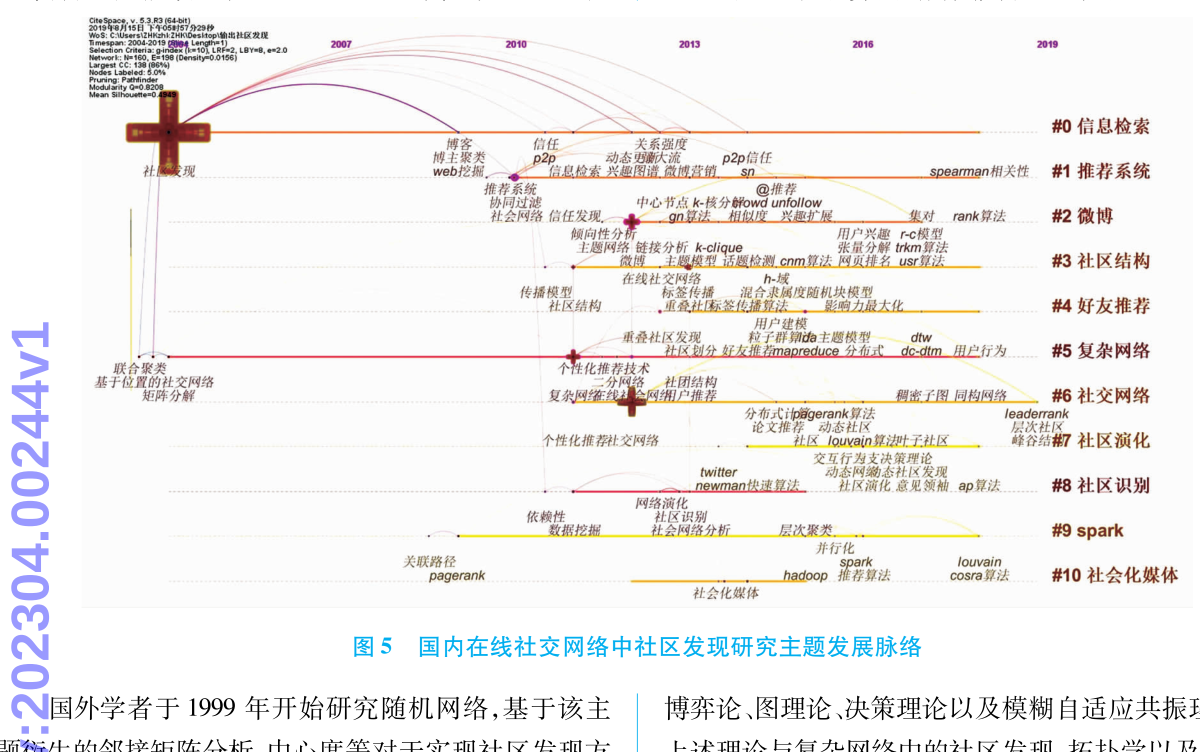


图 5　国内在线社交网络中社区发现研究主题发展脉络

国外学者于 1999 年开始研究随机网络，基于该主题衍生的邻接矩阵分析、中心度等对于实现社区发现方法至关重要（见图 6）。其他主题聚类也基本始于 2004 年，分析可知：国外研究实践性较强，大量算法和技术已经融入具体现实领域，如社交媒体、医学病毒、电信通信业以及电子商务等，倾向于将虚拟网络中的社区发现与现实世界的群体分析相结合，旨在应对大数据挑战并提供更好的用户体验。同时，国外更加注重理论研究，如博弈论、图理论、决策理论以及模糊自适应共振理论等，上述理论与复杂网络中的社区发现、拓扑学以及标签传播算法关联紧密，尤其是以博弈论为基础的社区发现方法在重叠社区识别中广受关注[7]。同样地，国外也重视网络社区结构识别、重叠社区检测以及动态社区演化的研究。在算法方面，国内外研究大体相似，图 6 中所展示的蝙蝠算法（BA 算法）是 X. S. Yang[8]于 2010 年提出的一种搜索全局最优解的有效方法。

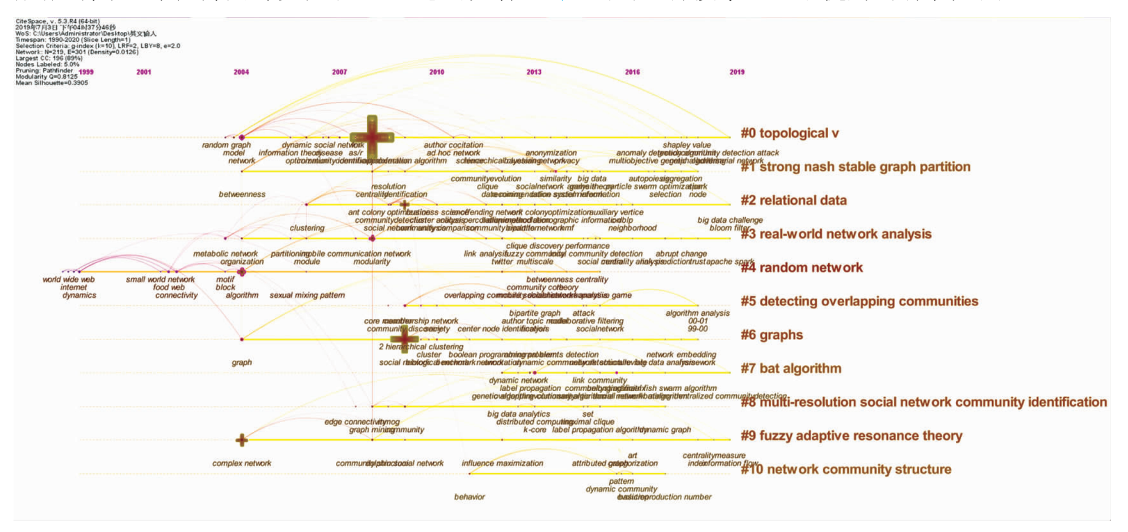


图 6　国外在线社交网络的社区发现研究主题发展脉络

2.5　关键文献分析

共被引文献分析呈现的节点文献在主题发展演化中起到承上启下作用，借助 CiteSpace 的突发性文献信息分析功能，同时逐年分析近 10 年的中文高被引文献，得出如图 7 和表 1 所示的代表学者及关键文献：

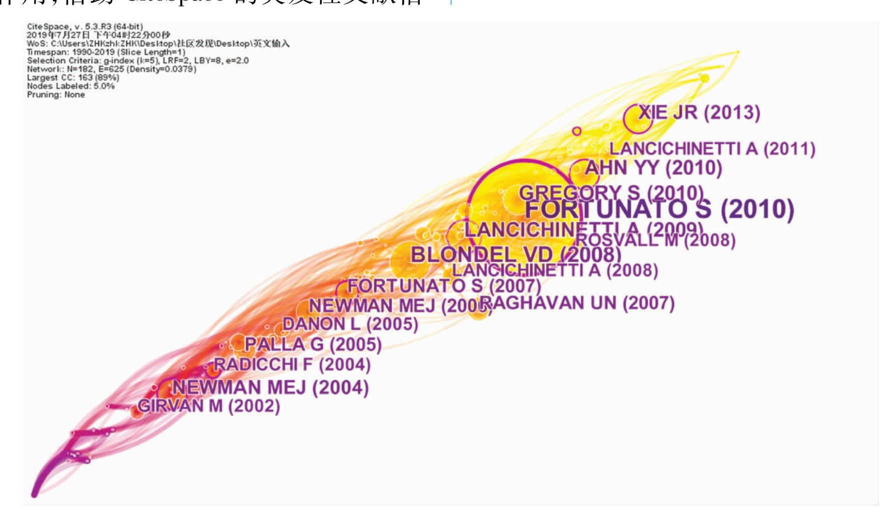


图 7　英文共被引文献分析

表 1　核心文献归纳

核心文献	代表学者	研究内容	时间(年)
[9－10]	M. Girvan, M. E. J. Newman 等	最早研究社交网络的社区结构，提出“模块度”这一重要概念	2002
[11]	F. Radicchi 等	按结构紧密程度将网络社区划分为强弱社区	2004
[12－14]	L. Danon，王莉，李建华等	综述：社区发现方法的对比分析及提升建议	2005－2015
[15－16]	U. N. Raghavan, R. Albert, M. Rosvall 等	较早提出了社区发现算法，引起学者关注	2004－2010
[17]	G. Palla, I. Derényi 等	首次定义重叠社区，成为新的研究重点	2005
[18]	Y. Y. Ahn 等	新的研究思路：认为社区是由一组密切相关的“链接”构成，而非节点	2010
[19－22]	S. Gregory，吴小兰，章成志，刘世超，辛宇等	社区发现算法的优化及发展	2010－2019
[23]	F. Radicchi, L. Danon, A. Lancichinetti 等	不断提出并改善算法的基准测试及评测标准	2007－2011
[24－26]	J. Xie, Z. Zhao，何婧等	探究动态社区发现算法，追踪社区结构演化	2013－2019

归纳上述关键文献可知：学术界更加重视重叠社区发现、动态社区识别以及社区发现方法的对比分析。同时基于在线社交网络近年来凸显的社区特性，社区发现算法及评价标准得到进一步关注和优化。由于下文针对社区发现方法及算法进行重点述评，本节仅简要归纳核心文献内容。

3　在线社交网络的社区发现主题内容分析

通过上文定量分析和文献梳理可知：该主题下社区发现方法、算法实现及应用实践方面的研究非常丰富，可以基本包含有关研究主题，还未有学者进行全面述评。因此，本文将在线社交网络的“社区发现”主题划分为社区发现的方法、算法实现以及实践应用研究 3 个维度进行定性述评。

3.1　社区发现方法

专家学者通过不同视角探究社区发现方法，紧跟网络的发展变化制定新的社区发现方案，包括多种理论或模型的结合及创新，本文将其概述为社区发现方法，通过各类方法研究，可以更好地处理数据集差异性问题并挖掘网络社区结构的细微之处。社区发现方法归纳见表 2。

3.1.1　基于图理论的社区发现方法(2002 年至今)

图是社交网络中常用的关系表现形式，包括节点、边和度等要素，经典的图聚类方法也称为图划分，指试图找到最好的切割方式将图划分为不同的部分(即社区)，常采用最小图分割方法。Y. R. Lin 等[27]研究媒体平台中的社区结构时，利用元图构建多关系模型和多维社交数据，通过增量元图分解来处理时变关系。J. Chen 等[28]从给定的稀疏图中提取有意义的稠密子图，文中“稠密子图”被作者解释为社区，该方法无需提前指定划分簇的数量。

表 2 社区发现方法归纳

方法	作者	研究内容	优缺点
基于图理论的社区发现方法	Y. R. Lin 等	通过分解增量元图处理时变关系的社区发现方法	理论发展完善、方法简单可行，但在识别动态网络时存在局限性
	J. Chen 等	从给定的稀疏图中提取有意义的稠密子图	
基于主题语义的社区发现方法	Z. Xia 等	利用评论内容挖掘语义信息，从而构建相似主题语义网络	可以借助关键语义信息，反映人们的偏好、关注话题等内容，适用性强
	M. M. Anwar 等	基于同一主题的输入查询，发现时间敏感型活动驱动的用户群体	
基于局部优化扩张的社区发现方法	K. Tang 等	提出局部影响优先策略，根据其边际增益为每个社区分配候选节点	不需要构建完整的网络拓扑结构，降低计算成本
基于数学模型的社区发现方法	P. K. Gopalan 等	基于贝叶斯网络模型，结合网络子采样动态更新社区的估计值	运算精确，识别的社区结构质量较高，但膨胀过程中存在参数过度敏感性问题
	张琴等	运用灰色理论、密度峰值聚类算法及粗糙集理论实现社区发现	
基于链接分析的社区发现方法	Y. Y. Ahn 等	认为社区是由"链接"构成，通过链接相似性进行社区识别	可以利用潜在链接有效检测社会群体行为的隐含联系，适用于大规模网络处理
	W. Liu 等	以马尔科夫网络为框架，利用链接分析网络中正在互动的群体	
基于深度学习算法的社区发现方法	L. Yang 等	基于模块度函数的半监督方法挖掘社区结构	高纬度的数据处理和数据挖掘能力，易于解决大型网络的复杂计算问题
	G. Sperli	利用深度学习算法改善邻接矩阵高维性或稀疏性	

3.1.2 基于数学模型的社区发现方法(2007 年至今)

该主题研究涉及大量数学领域的理论和方法，P. K. Gopalan[29]提出了一种基于贝叶斯模型的社区检测方法，允许节点参与多个社区，与重叠社区识别的特征相契合，同时灵活地叠加了来自网络的子采样并动态更新发现社区的估计值。张琴等[30]运用灰色理论定义全局结构相似性，结合密度峰值聚类算法确定聚类中心，并引入粗糙集理论根据网络结构自动选取中心节点，不断调整距离比率阈值进行划分迭代，从而划分重叠社区结构。

3.1.3 基于主题语义的社区发现方法(2009 年至今)

大量社区发现方法多采用节点共同属性或拓扑结构划分社区，无法利用节点或边缘的语义等关键信息，未能反映人们的兴趣爱好、关注话题等内容。基于主题语义的社区发现方法考虑了节点信息内容，适用于研究社交媒体平台。Z. Xia 等[31]挖掘评论内容中的语义信息构建整个语义主题网络，聚焦主题权重对每个边的影响，将重点放在降低计算复杂度上，适用于大规模网络处理。M. M. Anwar 等[32]针对一组给定查询主题，跟踪动态社交网络中时间敏感型驱动的用户群组，发现组内用户的主题关注情况在时间上倾向于相似。

3.1.4 基于链接分析的社区发现方法(2010 年至今)

社交媒体平台用户间关注存在单向情况，联系程度较弱，利用拓扑结构特性进行社区发现有时并不理想。该方法允许网络的顶点属于多个社区，有助于发现重叠社区结构。与大多数研究者观点不同，Y. Y. Ahn 等[18]指出高度重叠社区可能存在更多外部链接，认为社区是由一组密切相关的"链接"构成，而非节点，该方法使用层次结构聚类和链接之间的相似性构建树枝图，在最大分区密度处分析发现的链接社区。W. Liu 等[33]以马尔科夫网络为框架，通过分析与社交对象相关联的链接进行社区检测，并利用潜在链接挖掘社会群体行为中的隐含动态，该方法实现无需考虑社交网络的拓扑结构。

3.1.5 基于局部优化和扩张的社区发现方法(2012 年至今)

在大规模、具有众多节点信息的社交网络中识别虚拟社区时，局部社区发现算法不需要网络的整体信息，该类方法多通过局部结构信息快速定位目标节点所在社区，从核心节点出发，通过局部收益函数及贪婪策略将周围节点纳入已识别的虚拟社区中，主要包括局部扩展优化、派系过滤、标签传播、局部边聚类优化 4 类方法[14]。J. Tang 等[34]在实现社区传播影响力最大化时，采用了局部影响优先策略，在第一阶段利用标签传播算法检测社区分布，并根据其边际增益为每个社区分配候选节点数量，在此基础上制定了粒子个体的动态编码机制和群体离散演化规则，以此识别社区内高影响力节点。

3.1.6 基于深度学习算法的社区发现方法(2014 年至今)

在大数据时代，深度学习算法研究方兴未艾，广泛应用于计算机视觉、机器阅读理解以及大规模数据处理等领域。少数学者针对社区发现方法提出了解决方

案,其原理多为对网络的节点信息进行数据降维处理,或者通过训练网络图相似度矩阵得到低维特征矩阵。L. Yang 等[35]提出了一种基于模块度函数的半监督社区检测方法,同时使用未加权图的矩阵作为自动编码器输入,用于获得具有非线性映射的低维嵌入矩阵。G. Sperli[36]针对邻接矩阵的高维度或稀疏性问题,提出了一种基于深度学习的新型社区发现方法,充分考虑了数据集的维度和邻接矩阵的拓扑特性。

3.2 社区发现算法

算法研究对于挖掘社区结构至关重要,根据上文定量分析可知:社区发现算法的研究始于 2002 年,M. Girvan、M. E. J. Newman、S. Fortunato 等学者较早关注到该领域。2010 年以后,有关算法优化和创新的研究逐渐增多,专家学者逐步提出或通过优化早期社区发现算法来应对瞬息万变的社区结构。同时,算法的基准测试一直处于不断完善的状态,关于重叠社区和动态社区的探索迄今为止均为重点研究内容。因此,结合上文关键文献梳理,追踪最新算法研究,归纳出如下内容:

3.2.1 早期权威的社区发现算法

早期社区发现算法大体可以分为分裂式算法、基于模块度优化及信息论思想提出的相应算法。

(1) G-N 算法是最具代表性的分裂式算法,M. Girvan 和 M. E. J. Newman[9]于 2002 年聚焦社交网络和生物网络的社区结构,首次指出网络拓扑特征有助于社区结构识别,同时提出了基于最大边介数的分裂式层次算法(G-N 算法)。该算法不需要提前确定社区数目,在一定程度上改善了 Kernighan-Lin(KL)算法和谱二分法的局限性,但它需要重复评估每个边缘,成本较高。

(2) Louvain 算法[37]是一种基于模块度的社群发现算法,该算法无需事先明确群落信息,在研究大型网络上具有较好的效率和效果表现,常用于博客类社交媒体平台以及引文网络社区[38]。基于贪婪原理的 CNM 算法[39]本质上也应用了模块度的思想,该算法使用 3 个数据结构进行社区发现,其中稀疏矩阵即为表示节点模块度变化的变量。除此之外,基于模拟退火、极值优化提出的算法也源于模块度原理。

(3) M. Rosvall 等从信息论的视角出发,基于随机游走编码和信息压缩编码思想,于 2008 年提出了 Informa 算法[16]。该算法运用目标优化函数将相似度较大的节点分配到同一社区中,通过划分编码长度最短的虚拟社区从而达到社区发现的目的,多用于需要细致刻画社区结构的情形。

3.2.2 社区发现算法的发展及优化

以上算法假设每个节点仅存在于单个社区,然而以在线社交网络为例,每个用户都具备从属不同社区的可能性,因此,虽然有些算法沿用至今,但仍存在使用上的局限性。G. Pulla 等[17]首次证明社交网络中存在重复节点现象,提出了 Cluster Percolation Method (CPM)派系过滤算法。随着网络社区的结构特性受到广泛关注,出现了各类适用性更强的社区发现算法。

(1) 标签传播算法及其优化。标签传播算法 LPA (Label propagation algorithm)[40]是一种基于图的半监督学习方法,其思路是用已标记节点的标签信息去预测其他节点,具有实现简单以及时间复杂度低等优点,但 LPA 算法仅适用于非重叠的静态网络社区。S. Gregory[19]通过扩展节点标签类型及信息传播路径对标签传播算法做出优化,使信息值囊括多个社区,适用于重叠社区。吴小兰等[20]在深入研究多标签传播算法的基础上,利用节点对社区的贡献度来区分节点与其邻居社区的紧密程度,提出了基于贡献度的多标签传播算法 COPRA_CD。刘世超[21]提出一种基于标签传播概率的 LPPB 重叠社区发现算法,该算法首先为每个结点赋予一个独立标签,然后根据结点的影响力大小进行排序,综合网络传播特性和结点属性特征值来计算标签传播的概率,最后利用结点的历史标签记录修正发现结果。

(2) 基于谱分析和聚类的社区发现算法。基于谱分析的社区算法基于如下思想:相同社区内的节点在拉普拉斯矩阵中的特征向量呈近似性,将节点对应的矩阵特征向量视为空间坐标,将网络节点映射到多维向量空间中,然后运用 K-means 或 FCM 等经典算法聚集成社团,此类算法成本较大,但因其可以直接使用传统的向量聚类成果,灵活性很高[41]。T. Ma 等[42]针对重叠社区结构识别提出了 LED 算法,该算法基于结构聚类,将顶点间的结构相似性转换为网络权重,在算法精度及运算效率方面有所提升。张军祥[43]提出了一种基于平滑 L1 范数的深度稀疏自编码器社区发现算法 L1-ECDA,算法对网络图的邻接矩阵进行降维预处理,通过三层神经网络及 K-means 算法进行矩阵聚类得到隐藏的网络社区。

(3)融合多维信息的社区发现算法。用户交流逐渐转向网络媒体,社会关系也更多地出现于在线社交平台,针对平台特性的算法研究成为近年来的研究热点,算法通过融合多维信息提高其准确性,主要包括用户行为、节点内容、链接权重、社交关系及地理属性等。刘冰玉等[44]提出的DC-DTM算法将微博网络映射为有向加权网络,边的方向反映节点间关注关系,将节点间的语义相似度赋值为节点间连接权重,有效解决了微博网络的稀疏性以及传统LPA算法的逆流问题。田博等[45]提出了一种基于用户交互行为的社区发现算法,利用微博用户间的转发关系、评论关系以及提及关系等构建加权交互网络,不断合并使模块度函数增益最大的节点对,直到模块度函数增益为负。辛宇等[22]提出了一种面向语义社区发现的LBTC算法,该算法以LDA模型为语义信息模型,将语义特性和社会关系特性相结合,并通过定义语义链接权重实现了语义信息的可度量化。

(4)动态社区发现算法。社交网络的整体规模不断扩张,同时存在节点和链接消失现象,社区演化分析主要探究网络结构的阶段性变化,这种演变特征更能体现网络用户的动态交互行为。但学者大多将现实网络抽象为某一时间窗口的静态网络,算法仅适用于增加少量节点或边缘的情况。J. Xie等[24]提出的LabelRank算法对早期算法的随机性问题做出优化,引入一组操作符控制和稳定传播动态,从而识别出在网络中正在检测的所有社区,与静态社区发现算法相比,LabelRank算法显著提高了检测到的社区质量。Z. Zhao等[25]多年来致力于研究边缘权重的更新规则,提出"完全独立"子图更新策略,将动态社区按照新生社区的诞生以及原有社区的扩张两种方法划分,从而按时态挖掘社区演变过程。何婧等[26]提出的LUA算法考虑增量节点,依据拓扑势场理论更新拓扑势值,计算增量节点影响范围内的邻居节点,从而动态调整网络变化部分的社区归属。

3.2.3 社区发现算法的评测指标

(1)2004年,M. E. J. Newman和M. Girvan[10]提出了著名的"模块度"作为衡量社区划分的标准,采用G-N算法进行试验,模块度的提出为算法研究和算法评测带来新的契机,后期也有学者针对重叠社区发现算法提出了相应的模块度测评公式。

(2)F. Radicchi[11]首次以子图为依据将网络社区定量划分为"强社区和弱社区",可以对社区结构紧密性做出判断,在此基础上提出了仅考虑局部变量的分裂算法,对模块度优化存在的分辨率局限、过度依赖网络全局特征等问题做出改善。

(3)在算法的性能方面,L. Danon等[12]依据灵敏度和计算成本比较了当时的社区发现算法,认为在选择算法时应综合衡量这两个方面。除此之外,精度及时间复杂度也是衡量算法性能的常用指标。

(4)归一化互信息也是目前广泛使用的一种社区划分评价指标,该方法定义了一个混淆矩阵,其中行对应于真实社区,列对应于发现社区,通过度量发现社区和真实社区之间相似程度来判别算法的真实效果[12]。

(5)随着链路方向和算法权重受到广泛关注,A. Lancichinetti等[23]于2011年提出了一类无向和未加权的基准图,对节点度和社区规模分布的异质性做出阐释,描绘了节点度和社区的异构分布尺寸,并充分考虑节点属于多个社区的可能性,适用于重叠社区算法的测评。

综上可知,社区发现算法不断更新,对于算法的测试标准也在持续增加,但至今未能得到统一,尤其是动态社区内部存在合并、分裂、缩小、增大、产生和消失等各种情况[13],仅依据相邻时间点的社区相似度进行评价不够准确。其次,算法多通过合成网络或者自选测试网络进行验证,评测结果存在主观性。

3.3 社区发现应用实践

3.3.1 社交媒体平台中的应用

识别信息传播动向、发现社区结构演化模式和演化异常点,对于网络群体事件监测、舆情动态演进等具有重要价值。特别是实现动态社区识别的思路是基于时间节点的算法设计,与社会化媒体平台中的情境研究方向相契合。K. Gu等[46]将边缘紧密度和节点兴趣信息作为划分社区的参考标准,研究微博、YouTube和Digg社区的个人意愿和信息传播方案。C. Li等[47]提出了一种基于标签相关性和交互行为的微博社区发现算法,利用改进的最大边际相关性模型准确赋予用户标签,从而挖掘微博用户群体特性。李纲等[48]借鉴社区发现的思想,对构建的共词网络进行划分,形成描述不同热点话题的"话题社区",为微博热点话题的识别、测度和演化分析提供了新思路。研究发现,微博、Facebook以及Twitter等社交平台可以提供个人信息(节点属性),包括好友记录、兴趣偏好和位置信息等,

对其进行社区发现研究有助于及时掌握群体动向，提升用户体验。

3.3.2 推荐系统中的应用

社区发现方法可以在大规模用户群中识别相似用户，进而依据用户共同特征进行精准推送，尤其是局部社区发现和网络社区的链路预测都可以进一步优化传统协同过滤针对整个用户网络运算的数据量过载、推荐效率低等缺陷。K. Xinchang 等[49]通过构建用户关系网络缓解推荐系统中的冷启动问题，利用个人信息建立用户关系矩阵，并采用基于边缘中心性的社区检测方法为新用户提供精准推送服务。张继东等[50]通过构建基于社区划分和用户相似度的信息服务推荐模型，提高了移动社交网络中好友推荐效率及好友信息服务推荐的准确性和可信性。社区发现方法在保证推荐准度的前提下，大幅度提升推荐效率，为在线社交网络中开展精准推荐服务提供了新方法。

3.3.3 互联网营销中的应用

用户通过网络媒介建立社会联系，信息传播活动形成一种自然交互行为，伴随网络用户规模的不断扩大，互联网营销占据主流。但单纯在庞大的社交网络中进行推广营销，效率低且成本较大，因此可以利用社区发现方法提供解决方案：通过挖掘网络社区结构特征，有效地选择营销活动范围，避免信息传播重叠，提高网络服务效率。Y. C. Chen 等[51]基于影响力最大化问题开发了高影响力群体识别框架，根据社区规模分配种子数量，通过识别社区结构、选择候选人及确定种子节点 3 个步骤最大限度地扩大信息传播范围。因此，影响力最大化问题常指确定影响传播的最小节点集合，但对算法的效率和实用性要求较高。

4 研究难点及未来展望

研究发现：虽然算法类文献较多，但大多聚焦早期权威算法的优化和借鉴，早期算法在社区识别时出现的极大干扰情况未能做出准确说明，具体而言，分析动态社区演化时是否准确保留源历史信息？如何有效减少大型社区对弱社区的吞并现象而不是简单地判定为重叠社区？同时，G-N 算法的高成本问题、LPA 算法的不稳定性缺点以及其他算法高成本问题背后带来的回报率，如识别的准确性、识别重叠社区或者追踪动态社区演化等优势，应该通过何种统一标准权衡都亟待解决。其次，目前该领域的应用范围较窄，实际应用场景不够丰富，国内在线社交网络的研究平台目前局限于微博等内容分享类平台，应用领域多分布在虚拟社区中。

综上所述，当前较多算法不能准确反映真实网络社区现象，应结合在线社交网络的结构特性做出改善，并优化社区发现算法的评价指标。其次，移动互联网高速发展，5G 智能应用也即将面世，现实空间与虚拟网络的融合将不断深入，因此结合场景的社区发现方法值得重点关注。最后，该主题是在众多学科影响下发展起来的，涉及数学、物理、生物、计算机以及图书情报等领域，未来在进行社区发现的相关研究时，结合不同学科的研究方法和工具可能会成为解决问题的突破口。因此，结合当前研究难点，提出如下未来展望：

4.1 社区发现算法的优化及创新

现有算法大多高度复杂，不利于在大规模社交网络中挖掘网络社区，会造成一定程度的信息缺失。其次，网络社区结构更新迅速，采用大规模静态数据计量的算法不再适用。最后，关注到用户从内容创意型社交网络逐步向消息流型社交网络迁移，基于位置的发现算法未能体现社交媒体中的时序性特征。因此，未来在算法方面可以开展以下研究：

（1）针对快速算法、模糊识别算法的研究，例如基于局部网络节点、关键标签信息、融合多维节点信息的社区发现算法在未来应受到进一步重视。

（2）优化早期权威算法适应性问题，构建社区发现算法的评价机制，统一社区发现算法的评判标准，给予不同指标相应权值，尤其要完善动态社区发现算法的评价问题，应对不同时间窗口的社区演化和识别的社区质量进行综合评价。

（3）提出基于时空特征的社区发现算法，结合时间和空间信息对基于位置的发现算法做出优化，研究社交媒体中话题或者用户群体的时空轨迹分布情况。

4.2 区分并拓展社区发现方法的应用场景

（1）区分应用场景，选择最佳方法。例如较小的社群分析，可以选择运算准确但效率偏低的算法；相反，在研究具有千万级节点的网络时，除识别准确性以外，还应考虑预算和时间成本问题，在社区发现方法的选择上不能盲目追求其识别的精准性。

（2）扩展研究平台类型。有关淘宝、亚马逊等购物平台、在线健康社区以及知乎为代表的在线问答社区研究较少，可以依托各类平台开展社区发现研究。

(3)拓宽实际应用场景。国外较多文献涉猎于病毒传播网络、通信网络以及网络犯罪团体识别等相关领域。未来可以拓展社区发现实际应用场景,如社区结构分析中的用户隐私保护、发现异常用户群体解决网络安全问题以及虚拟学术社区中研究新兴主题识别等。

4.3 融合跨学科知识、前沿技术方法的交叉研究

(1)融合跨学科领域知识。例如借鉴网络结构分析方法在研究蛋白质网络结构、预测蛋白质病变以及分子间相互影响的案例[52],未来可以融合医学及生物学领域等知识,结合社区发现方法分析生物细胞、分子结构等。或基于复杂网络与传播动力学理论,利用社区发现方法研究信息和知识的传播机理、学科间的知识融合规律及追踪网络舆情传播态势等,都是值得关注的课题。

(2)结合人工智能类的新兴方法。复杂网络的节点信息表示多采用人工提取特征的方法,网络表示学习算法将网络信息转化为低维稠密的实数向量,通过梯度下降优化算法实现最优化求解参数,可以降低计算量和人工成本[53]。如SOM神经网络聚类算法[54]可以赋予不同标签相应权重值,与社区发现方法相结合可以对用户进行准确聚类,有效解决兴趣特征的稀疏性或特征选择过程中标签过多造成的维数灾难。其次,人们对视频媒体平台逐步热衷,开展视频媒体中的语义信息提取、聚焦用户生成内容等均为难点所在,导致社区发现方法在图文、视频分享类媒体中研究较少,以word2vec词向量生成模型、卷积神经网络在图像识别、自然语言处理上的成功应用为参考,未来可以将前沿的技术方法应用于社区发现方法的提升中,有助于在基于内容的多媒体社区中挖掘群体结构。

参考文献:

[1] 方滨兴,贾焰,韩毅.社交网络分析核心科学问题、研究现状及未来展望[J].中国科学院院刊,2015,30(2): 187-199.

[2] 程学旗,沈华伟.复杂网络的社区结构[J].复杂系统与复杂性科学,2011,8(1): 57-70.

[3] 吴小兰,章成志.学术社交媒体视角下学科知识流动规律研究——以科学网为例[J].数据分析与知识发现,2019,3(4): 107-116.

[4] 舒文琛,周恩国,李岱峰,等.基于合著网络社区发现的情报学研究主题演化分析[J].情报科学,2020,38(1):75-81.

[5] NEWNAN M E J. Networks[M]. New York: Oxford University Press, 2018.

[6] 杨楠,弓丹志,李忺,等.Web社区发现技术综述[J].计算机研究与发展,2005(3):439-447.

[7] ZHOU L, LU K, YANG P, et al. An approach for overlapping and hierarchic al community detection in social networks based on coalition formation game theory[J]. Expert systems with applications, 2015, 42(24):9634-9646.

[8] YANG X S. A new metaheuristic bat-inspired algorithm[C]//Nature inspired cooperative strategies for optimization. Berlin: Springer-Verlag,2010: 65-74.

[9] GIRVAN M, NEWNAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[10] NEWNAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 1-15.

[11] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences, 2004, 101(9): 2658-2663.

[12] DANON L, DIAZ-GUILERA A, DUCH J, et al. Comparing community structure identification[J]. Journal of statistical mechanics: theory and experiment, 2005, 2005(9): 1-10.

[13] 王莉,程学旗.在线社会网络的动态社区发现及演化[J].计算机学报,2015,38(2):219-237.

[14] 李建华,汪晓锋,吴鹏.基于局部优化的社区发现方法研究现状[J].中国科学院院刊,2015,30(2):238-247,180.

[15] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 1-11.

[16] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the national academy of sciences, 2008, 105(4): 1118-1123.

[17] PALLA G, DERÉNYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043):814-818.

[18] AHN Y Y, BAGROW J P, LEHMANN S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466(7307): 761-764.

[19] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New journal of physics, 2010, 12(10): 1-26.

[20] 吴小兰,章成志.基于贡献度的多标签传播重叠社区发现研究[J].情报学报,2015,34(9):949-957.

[21] 刘世超,朱福喜,甘琳.基于标签传播概率的重叠社区发现算法[J].计算机学报,2016,39(4):717-729.

[22] 辛宇,杨静,谢志强.一种面向语义重叠社区发现的Link-Block算法[J].软件学报,2016,27(2):363-380.

[23] LANCICHINETTI A, RADICCHI F, RAMASCO J J, et al. Finding statistically significant communities in networks[J]. PloS one,

2011, 6(4): 1－18.

[24] XIE J, SZYMANSKI B K. LabelRank: a stabilized label propagation algorithm for community detection in networks[C]// The 2013 IEEE 2nd international network science workshop. New York: IEEE, 2013: 138－143.

[25] ZHAO Z, LI C, ZHANG X, et al. An incremental method to detect communities in dynamic evolving social networks[J]. Knowledge-based systems, 2019, 163: 404－415.

[26] 何婧,王志晓,侯梦男,等.基于拓扑势的增量式动态社区发现方法[J].计算机工程与设计,2019,40(1):45－52.

[27] LIN Y R, SUN J, CASTRO P, et al. Metafac: community discovery via relational hypergraph factorization[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. New York : ACM, 2009: 527－536.

[28] CHEN J, SAAD Y. Dense subgraph extraction with application to community detection[J]. IEEE transactions on knowledge and data engineering, 2010, 24(7): 1216－1230.

[29] GOPALAN P K, BLEI D M. Efficient discovery of overlapping communities in massive networks[J]. Proceedings of the National Academy of Sciences, 2013, 110(36): 14534－14539.

[30] 张琴,陈红梅,封云飞.一种基于粗糙集和密度峰值的重叠社区发现方法[J/OL].[2020－02－01]. http://kns.cnki.net/kcms/detail/50.1075.tp.20200106.0943.007.html.

[31] XIA Z, BU Z. Community detection based on a semantic network[J]. Knowledge-based systems, 2012, 26: 30－39.

[32] ANWAR M M, LIU C, LI J. Discovering and tracking query oriented active online social groups in dynamic information network[J]. World Wide Web, 2019, 22(4): 1819－1854.

[33] LIU W, YUE K, WU H, et al. Markov-network based latent link analysis for community detection in social behavioral interactions[J]. Applied intelligence, 2018, 48(8): 2081－2096.

[34] TANG J, ZHANG R, YAO Y, et al. An adaptive discrete particle swarm optimization for influence maximization based on network community structure[J]. International journal of modern physics C, 2019, 30(6): 1－21.

[35] YANG L, CAO X, HE D, et al. Modularity based community detection with deep learning[C]//Proceedings of the Twenty-Fifth international joint conference on artificial intelligence (IJCAI-16). New York: AAAI Press. 2016: 2252－2258.

[36] SPERLI G. A deep learning based community detection approach[C]// Association for Computing Machinery. Proceedings of the 34th ACM/SIGAPP symposium on applied computing. New York : ACM, 2019: 1107－1110.

[37] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008,2008(10): P10008.

[38] 张海涛,刘雅姝,张枭慧,等.基于模块度的话题发现及网民情感波动研究——以新浪微博“中美间贸易摩擦”话题为例[J].图书情报工作,2019,63(4):6－14.

[39] CLAUSET A, NEWMAN M E J, MOORE C. Finding community structure in very large networks[J]. Physical review E, 2004, 70(6): 066111.

[40] ZHU X J, GHAHRAMANI Z, LAFFERTY J. Semi-supervised learning using Gaussian fields and harmonic functions[C]//Proceedings of the 20th international conference on machine learning. Washington DC: ICML. 2003:912－919.

[41] 邢翔瑞. Graph 特征提取方法:谱聚类(Spectral Clustering)详解[EB/OL].[2019－11－10]. https://blog.csdn.net/weixin_36474809/article/details/89669623? utm_source = app.

[42] MA T, WANG Y, TANG M, et al. LED: a fast overlapping communities detection algorithm based on structural clustering[J]. Neurocomputing, 2016, 207: 488－500.

[43] 张军祥,李书琴,刘斌.基于平滑 L1 范数的深度稀疏自动编码器社区识别算法[J/OL].[2020－02－05]. https://doi.org/10.19734/j.issn.1001－3695.2018.09.0743.

[44] 刘冰玉,王翠荣,王聪,等.基于动态主题模型融合多维数据的微博社区发现算法[J].软件学报,2017,28(2):246－261.

[45] 田博,凡玲玲.基于交互行为的在线社会网络社区发现方法研究[J].情报杂志,2016,35(11):183－188.

[46] GU K, WANG L, YIN B. Social community detection and message propagation scheme based on personal willingness in social network[J]. Soft computing, 2019, 23(15): 6267－6285.

[47] LI C, BAI J, DU S, et al. Combining tag correlation and interactive behaviors for community discovery[J]. The computer journal, 2018, 62(5): 785－800.

[48] 李纲,陈思菁,毛进,等.自然灾害事件微博热点话题的时空对比分析[J].数据分析与知识发现,2019,3(11):1－15.

[49] XINCHANG K, VILAKONE P, PARK D S. Movie recommendation algorithm using social network analysis to alleviate cold-start problem[J]. Journal of information processing systems, 2019, 15(3): 616－631.

[50] 张继东,蔡雪.基于社区划分和用户相似度的好友信息服务推荐研究[J].情报理论与实践,2019,42(4):151－157,165.

[51] CHEN Y C, ZHU W Y, PENG W C, et al. CIM: Community-based influence maximization in social networks[J]. ACM transactions on intelligent systems and technology, 2014, 5(2):1－31.

[52] KOVÁCS I A, LUCK K, SPIROHN K, et al. Network-based prediction of protein interactions[J]. Nature communications, 2019, 10(1):1－8.

[53] 涂存超,杨成,刘知远,等.网络表示学习综述[J].中国科学:信息科学,2017,47(8):980－996.

[54] 丁永刚,张雨琴,付强,等.基于 SOM 神经网络和排序因子分解机的图书资源精准推荐[J].情报理论与实践,2019,42(9):133－138,170.

作者贡献说明:
张海涛:研究方向和基本框架提出,论文指导及修改;
周红磊:数据采集,可视化分析处理,论文撰写;
张鑫蕊:文献整理;
孙彤:论文修订。

Research Progress in Community Detection of Online Social Networks

Zhang Haitao[1,2] Zhou Honglei[1] Zhang Xinrui[1] Sun Tong[1]
[1] Management School of Jilin University, Changchun 130022
[2] The Information Resource Research Center of Jilin University, Changchun 130022

Abstract: [Purpose/significance] Taking online social network as the research object, and through the literature combing to accurately capture the development trend and research hotspots of community discovery, and exploring how to mine hidden communities in large-scale social networks, which has theoretical and practical significance. [Method/process] Using CNKI database, Web of Science core collection and related international conference documents as data sources. The CiteSpace visual analysis tool was used to quantitatively study hotspot keywords, topic evolution paths and co-cited documents. And the topic research content was reviewed from 3 dimensions: community discovery method, algorithm implementation and application practice. [Result/conclusion] There is still much room for development in the current research field. In the future, we should pay attention to optimization and innovation of algorithms, differentiation and expansion of application scenarios, and cross-disciplinary research on interdisciplinary knowledge and cutting-edge technology methods.

Keywords: online social network community detection dynamic community evolution research progress